International Journal of Mathematical Education in Science and Technology, Vol. 44, No. 3, 2013, 410–417
(author's reprint)

On equivalent characterizations of convexity of functions

Eleftherios Gkioulekas*

University of Texas-Pan American, Edinburg, TX 78539-2999
(*Received 04 Jan 2012, Accepted 24 Sep 2012, Published online: 19 Nov 2012*)

A detailed development of the theory of convex functions, not often found in complete form in most textbooks, is given. We adopt the strict secant line definition as the definitive definition of convexity. We then show that for differentiable functions, this definition becomes logically equivalent with the first derivative monotonicity definition and the tangent line definition. Consequently, for differentiable functions, all three characterizations are logically equivalent.

1. Introduction

The convexity of functions and the connection between convexity and the second derivative is an important topic in Calculus that is often undertreated by most mainstream Calculus textbooks. In general, for the case of differentiable functions, there are three possible ways to define a *convex up* function and a *convex down* function: there is the *tangent line definition*, the *first derivative monotonicity definition*, and the *secant line definition*. For a convex up function, the tangent line definition states that the graph of the function lies above every tangent line to the function for all contact points within a specified interval. The first derivative monotonicity definition states that the first derivative of the function is strictly increasing within the given interval. The secant line definition states that for any two points on the graph of the function within the given interval, the line segment connecting the two points is always above the graph of the function within the subinterval defined by the chosen two points.

Most mainstream calculus textbooks fail to explain that for differentiable functions all three definitions are rigorously equivalent to each other, and that the definitive definition of convexity should be the strict secant line definition, as it is the only definition that does not require the function to be differentiable. For example, Stewart [1] uses the tangent line definition, and tucked away in an appendix, away from prying student eyes, includes a proof that the first derivative monotonicity definition implies the tangent line definition, but does not bother to prove the converse statement, and completely disregards the secant line definition. Thomas [2] uses the first derivative monotonicity definition to define convexity, since its connection to the second derivative theorem is trivially obvious, and completely disregards the other two definitions. Even Varburg [3], which in many other regards is a more rigorous textbook than both Stewart [1] and Thomas [2], also takes the easy way out, and mentions only the first derivative monotonicity definition. Lang's Calculus textbook [4] treats convexity in a very non-rigorous manner by calling it "bending up" or "bending down", not even using standard terminology. More rigorous texts, such as Lang's analysis textbook [5] and Apostol's Calculus textbook [6], correctly use the secant line definition to define convexity and show

*Corresponding author. Email: gkioulekase@utpa.edu

that if the first derivative is increasing then the function is convex up, and likewise for the convex down case. Even these books, however, do not prove the converse statement or establish the equivalence of the secant line definition with the tangent line definition. Furthermore, whereas Lang [5] at least distinguishes between weak and strict convexity, Apostol's treatment [6] is unnecessarily limited to weak convexity.

The goal of this paper is to give a detailed and complete treatment of the theory of convexity. We will begin by stating the strict version of the secant line definition as the definitive definition of convexity. Then, we will show that for differentiable functions, the secant line definition is equivalent to the other two definitions. Given these results, for twice differentiable functions, it is a trivial corollary to establish the connection between the second derivative and convexity.

2. The main argument

Let $f : A \to \mathbb{R}$ be a function with $A \subseteq \mathbb{R}$, and let $[x_1, x_2] \subseteq A$ be an interval within the domain A of the function. The strict secant line definition of convexity reads:

$$\begin{aligned} f \text{ convex up at } [x_1,x_2] &\iff \forall a \in [x_1,x_2) : \forall b \in (a,x_2] : \forall t \in (0,1) : \\ &\quad : f(a+t(b-a)) < (1-t)f(a) + tf(b), \\ f \text{ convex down at } [x_1,x_2] &\iff \forall a \in [x_1,x_2) : \forall b \in (a,x_2] : \forall t \in (0,1) : \\ &\quad : f(a+t(b-a)) > (1-t)f(a) + tf(b). \end{aligned}$$

To give the geometric motivation for this definition, we first note that for $t \in (0,1)$, $x = a + t(b-a)$ gives all the points with x coordinate between $x = a$ and $x = b$. We also note that the equation of the secant line connecting the points $(a, f(a))$ and $(b, f(b))$ is given by

$$\begin{aligned} (\ell) : y &= f(a) + \frac{f(b)-f(a)}{b-a}(x-a) = f(a) + \frac{f(b)-f(a)}{b-a}t(b-a) \\ &= f(a) + t(f(b)-f(a)) = (1-t)f(a) + f(b). \end{aligned}$$

Since the convex up definition requires that the secant line (ℓ) should be above the graph of the function for any point between $x = a$ and $x = b$ (i.e. for $x = a+t(b-a)$ with $t \in [0,1]$), we obtain the inequality $f(a + t(b-a)) < (1-t)f(a) + tf(b)$ appearing in the definition above. Obviously, to maintain convexity, this inequality has to hold for any choice $a, b \in [x_1, x_2]$ with $a < b$, which can be enforced by restricting a to $[x_1, x_2)$ and b to $(a, x_2]$. Note that for $a = x_2$, it is impossible to choose an $b > a$ within the interval $[x_1, x_2]$, which is why we exclude x_2 from the possible values of a. The need for these restrictions can be eliminated if one opts for using *weak* convexity instead, which can be defined more succinctly as

$$\begin{aligned} f \text{ weakly convex up at } [x_1,x_2] &\iff \forall a,b \in [x_1,x_2] : \forall t \in (0,1) : \\ &\quad : f(a+t(b-a)) \leq (1-t)f(a) + tf(b), \\ f \text{ weakly convex down at } [x_1,x_2] &\iff \forall a,b \in [x_1,x_2] : \forall t \in (0,1) : \\ &\quad : f(a+t(b-a)) \geq (1-t)f(a) + tf(b). \end{aligned}$$

It may be tempting to believe that one has to settle for weak forms of the convexity definitions, where weak inequalities are used everywhere, in order to be able to show that they are all logically equivalent with each other. It is well-known, for example, that for the case of function monotonicity, a strictly positive first derivative over an interval implies that the function is strictly increasing over the same interval, but that the converse statement is not true, unless we formulate a more complicated theorem or settle for the weak monotonicity definitions and weak inequalities for the first derivative. One of the main points that we wish to stress in this paper is that the strong versions of all characterizations of convexity can indeed be shown to be equivalent with each other for differentiable functions. After all, it is the strict convexity definition that is more consistent with our intuitive mental image of a convex up or down function.

Now, given $a, b \in [x_1, x_2]$ with $a < b$, let us define the slope $\lambda(a, b)$ of the corresponding secant line from $x = a$ to $x = b$ as

$$\lambda(a,b) = \frac{f(b) - f(a)}{b - a}.$$

The first step of the argument is to prove the following lemma:

Lemma 2.1:

(1) If f is convex up at $[x_1, x_2]$ and $a, b, c \in [x_1, x_2]$ and $a < c < b$, then $\lambda(a, c) < \lambda(a, b) < \lambda(c, b)$.
(2) If f is convex down at $[x_1, x_2]$ and $a, b, c \in [x_1, x_2]$ and $a < c < b$, then $\lambda(a, c) > \lambda(a, b) > \lambda(c, b)$.

Proof: Assume, with no loss of generality, that f is convex up at $[x_1, x_2]$. Let $a, b, c \in [x_1, x_2]$ be given with $a < c < b$. Let $t \in (0, 1)$ such that $c = a + t(b - a)$. Then we note that

$$\begin{aligned} c - a &= a + t(b - a) - a = t(b - a), \\ b - c &= b - a - t(b - a) = (1 - t)(b - a). \end{aligned}$$

It follows that

$$\begin{aligned} \lambda(a,c) &= \frac{f(c) - f(a)}{c - a} = \frac{f(a + t(b-a)) - f(a)}{t(b-a)} \\ &< \frac{(1-t)f(a) + tf(b) - f(a)}{t(b-a)} = \frac{tf(b) - tf(a)}{t(b-a)} \\ &= \frac{f(b) - f(a)}{b - a} = \lambda(a,b) \Longrightarrow \lambda(a,c) < \lambda(a,b), \end{aligned}$$

and

$$\begin{aligned} \lambda(c,b) &= \frac{f(b) - f(c)}{b - c} = \frac{f(b) - f(a + t(b-a))}{(1-t)(b-a)} \\ &> \frac{f(b) - (1-t)f(a) - tf(b)}{(1-t)(b-a)} = \frac{(1-t)(f(b) - f(a))}{(1-t)(b-a)} \\ &= \frac{f(b) - f(a)}{b - a} = \lambda(a,b) \Longrightarrow \lambda(a,b) < \lambda(c,b). \end{aligned}$$

□

The geometric interpretation of this lemma is that given any secant line within an interval $[x_1, x_2]$ where your function is, for example, convex up, if you "slide" either endpoint of the secant line forward, then the slope of the secant line will increase.

For the argument below, we need to recall the definition for strict monotonicity. Given a function $f : A \to \mathbb{R}$ with $I \subseteq A$ an interval, we define:

$$f \uparrow I \Longleftrightarrow \forall x_1, x_2 \in I : (x_1 < x_2 \Longrightarrow f(x_1) < f(x_2)),$$
$$f \downarrow I \Longleftrightarrow \forall x_1, x_2 \in I : (x_1 < x_2 \Longrightarrow f(x_1) > f(x_2)).$$

Here $f \uparrow I$ reads "f is strictly increasing in I" and $f \downarrow I$ reads "f is strictly decreasing in I".

We also recall that given two functions $f : A \to \mathbb{R}$ and $g : A \to \mathbb{R}$ with $N(\sigma, \delta)$ a generalized neighborhood of the limit $x \to \sigma$, where σ is a generalized accumulation point of A (e.g. $\sigma = x_0$, or $\sigma = x_0^+$, or $\sigma = x_0^-$ with $x_0 \in \mathbb{R}$, or $\sigma = +\infty$ or $\sigma = -\infty$), it can be shown that:

$$(\exists \delta > 0 : \forall x \in N(\sigma, \delta) : f(x) < g(x)) \Longrightarrow \lim_{x \to \sigma} f(x) \leq \lim_{x \to \sigma} g(x).$$

It should be stressed that σ is a generalized accumulation point and $N(\sigma, \delta)$ is the corresponding generalized neighborhood, defined as

$$N(\sigma, \delta) = \begin{cases} (x_0 - \delta, x_0 + \delta) - \{x_0\} & , \text{ if } \sigma = x_0 \\ (x_0 - \delta, x_0) & , \text{ if } \sigma = x_0^- \\ (x_0, x_0 + \delta) & , \text{ if } \sigma = x_0^+ \\ (1/\delta, +\infty) & , \text{ if } \sigma = +\infty \\ (-\infty, -1/\delta) & , \text{ if } \sigma = -\infty \end{cases}.$$

Thus the statement above holds for regular limits, side limits, and limits to infinity, all of which can be subsumed under one proof via the generalized neighborhood $N(\sigma, \delta)$.

The point here is that given a strict inequality between two functions, taking the limit on both sides of the inequality reduces it to a weak inequality. The counterexample where $f(x) = -1/x^2$ and $g(x) = +1/x^2$ with $x \to +\infty$ shows that the above statement cannot be strengthened to give a strict equality on the right hand side of the implication. This complicates the argument below, but in spite of that we can still retain strict inequalities throughout all characterizations of convexity.

Taking the above preliminaries into consideration, and using Lemma 2.1, we will now show the equivalence between the strict secant line definition and the first derivative monotonicity definition:

Theorem 2.2: *Assume that f is differentiable on $[x_1, x_2]$. Then:*

$$f \text{ convex up at } [x_1, x_2] \Longleftrightarrow f' \uparrow [x_1, x_2],$$
$$f \text{ convex down at } [x_1, x_2] \Longleftrightarrow f' \downarrow [x_1, x_2].$$

Proof: ($\Longrightarrow$): Assume, with no loss of generality, that f is convex up at $[x_1, x_2]$. Let $a, b \in [x_1, x_2]$ be given with $a < b$. Let $h \in (0, \delta)$ with $\delta > 0$ chosen such that $a + \delta < b - \delta$ and $a + \delta < (a + b)/2$ and $b - \delta > (a + b)/2$. Then, via Lemma 2.1, we have:

$$\begin{aligned}\forall h \in (0, \delta) : a < a + h < (a + b)/2 &\Longrightarrow \forall h \in (0, \delta) : \lambda(a, a + h) < \lambda(a, (a + b)/2)\\ &\Longrightarrow f'(a) = \lim_{h\to 0^+} \lambda(a, a + h) \le \lambda(a, (a + b)/2),\end{aligned}$$

and similarly we have

$$\begin{aligned}\forall h \in (0, \delta) : (a + b)/2 < b - h < b &\Longrightarrow \forall h \in (0, \delta) : \lambda(b - h, b) > \lambda((a + b)/2, b)\\ &\Longrightarrow f'(b) = \lim_{h\to 0^+} \lambda(b - h, b) \ge \lambda((a + b)/2, b).\end{aligned}$$

The existence of the limits above is ensured by the differentiability of f. Combining the equations above, and with another application of the lemma, we have:

$$f'(a) \le \lambda(a, (a + b)/2) < \lambda((a + b)/2, b) \le f'(b),$$

and therefore $f'(a) < f'(b)$. It follows that $f \uparrow [x_1, x_2]$

($\Longleftarrow$): Assume now, with no loss of generality, that $f' \uparrow [x_1, x_2]$. Let $a, b \in [x_1, x_2]$ and $t \in (0, 1)$ be given with $a < b$. It follows that $a \in [x_1, x_2)$ and $b \in (a, x_2]$. Define $c = a + t(b - a)$, and note that $c - a = t(b - a)$ and $b - c = (1 - t)(b - a)$, therefore $a < c < b$. Since f is differentiable on $[x_1, x_2]$, we apply the mean-value theorem on the intervals $[a, c]$ and $[b, c]$, and obtain

$$\begin{aligned}\exists \xi_1 \in (a, c) :& f(c) - f(a) = f'(\xi_1)(c - a) = t(b - a)f'(\xi_1),\\ \exists \xi_2 \in (c, b) :& f(b) - f(c) = f'(\xi_2)(b - c) = (1 - t)(b - a)f'(\xi_2).\end{aligned}$$

Note that $\xi_1 < c < \xi_2 \Longrightarrow \xi_1 < \xi_2 \Longrightarrow f'(\xi_1) < f'(\xi_2)$, and also that:

$$\begin{aligned}f(c) - (1 - t)f(a) - tf(b) &= [f(c) - f(a)] - t[f(b) - f(a)]\\ &= [f(c) - f(a)] - t[f(b) - f(c)] - t[f(c) - f(a)]\\ &= t(b - a)f'(\xi_1) - t(1 - t)(b - a)f'(\xi_2) - t^2(b - a)f'(\xi_1)\\ &= t(1 - t)(b - a)f'(\xi_1) - t(1 - t)(b - a)f'(\xi_2)\\ &= t(1 - t)(b - a)[f'(\xi_1) - f'(\xi_2)].\end{aligned}$$

Since $t(1 - t) > 0$ and $b - a > 0$ and $f'(\xi_1) - f'(\xi_2) < 0$, we have $t(1 - t)(b - a)[f'(\xi_1) - f'(\xi_2)] < 0$, and therefore $f(c) < (1 - t)f(a) + tf(b)$. We have thus shown that

$$\forall a \in [x_1, x_2) : \forall b \in (a, x_2] : \forall t \in (0, 1) : f(a + t(b - a)) < (1 - t)f(a) + tf(b),$$

and therefore that f is convex up at $[x_1, x_2]$. □

The proof of the converse statement in the theorem above can be found in various forms in some textbooks like Apostol [6] and Stewart [1] (albeit, sequestered in an appendix). However, the forward part of the proof is not easy to find in textbooks and I had to derive it on my own. To ensure the restoration of the strict inequality it was necessary to employ a clever trick whereby Lemma 2.1 is applied multiple times.

The next step is to establish the equivalence between the first derivative monotonicity definition and the tangent line definition. Given the function f, let us define $g(x|x_0)$ as

$$g(x|x_0) = f(x) - [f'(x_0)(x - x_0) + f(x_0)].$$

The obvious interpretation of $g(x|x_0)$ is that it measures the difference in y-coordinates between a point $(x, f(x))$ on the graph of the function f and a point with the same x coordinate on the tangent line of the function f with contact point chosen at $(x_0, f(x_0))$. It follows that $g(x|x_0) > 0, \forall x \in [x_1, x_2] - \{x_0\}$ implies that the graph of the function f is *above* the given tangent line, whereas $g(x|x_0) < 0, \forall x \in [x_1, x_2] - \{x_0\}$ implies that the graph of the function f is instead *below* the given tangent line. Consequently, we have to prove the following theorem:

Theorem 2.3: *Assume that f is differentiable at $[x_1, x_2]$. Then:*

$$\begin{aligned} f' \uparrow [x_1, x_2] &\Longleftrightarrow \forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) > 0), \\ f' \downarrow [x_1, x_2] &\Longleftrightarrow \forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) < 0). \end{aligned}$$

Proof: ($\Longrightarrow$): Assume, with no loss of generality, that $f' \uparrow [x_1, x_2]$. Let $x, x_0 \in [x_1, x_2]$ be given such that $x \neq x_0$. We note that

$$g'(x|x_0) = (\mathrm{d}/\mathrm{d}x)[f(x) - f'(x_0)(x - x_0) - f(x_0)] = f'(x) - f'(x_0).$$

It follows that

$$\begin{aligned} \forall x \in (x_0, x_2) : x > x_0 &\Longrightarrow \forall x \in (x_0, x_2) : f'(x) > f'(x_0) \\ &\Longrightarrow \forall x \in (x_0, x_2) : g'(x|x_0) > 0 \\ &\Longrightarrow g(x|x_0) \uparrow (x_0, x_2) \text{ with respect to } x \\ &\Longrightarrow \forall x \in (x_0, x_2) : g(x|x_0) > g(x_0|x_0) = 0, \end{aligned}$$

and

$$\begin{aligned} \forall x \in (x_1, x_0) : x < x_0 &\Longrightarrow \forall x \in (x_1, x_0) : f'(x) < f'(x_0) \\ &\Longrightarrow \forall x \in (x_1, x_0) : g'(x|x_0) < 0 \\ &\Longrightarrow g(x|x_0) \downarrow (x_1, x_0) \text{ with respect to } x \\ &\Longrightarrow \forall x \in (x_1, x_0) : g(x|x_0) > g(x_0|x_0) = 0, \end{aligned}$$

and therefore: $\forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) > 0)$.

($\Longleftarrow$): Assume, with no loss of generality, that $\forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) > 0)$. Let $a, b \in [x_1, x_2]$ be given such that $a < b$. From the hypothesis, we have:

$$\begin{aligned} g(a|b) > 0 &\Longrightarrow f(a) - [f'(b)(a - b) + f(b)] > 0 \\ &\Longrightarrow f'(b)(a - b) < f(a) - f(b), \end{aligned}$$

and

$$\begin{aligned} g(b|a) > 0 &\Longrightarrow f(b) - [f'(a)(b - a) + f(a)] > 0 \\ &\Longrightarrow f'(a)(b - a) < f(b) - f(a). \end{aligned}$$

It follows that

$$[f'(a) - f'(b)](b-a) = f'(a)(b-a) + f'(b)(a-b) < \\ < [f(b) - f(a)] + [f(a) - f(b)] = 0,$$

and therefore $f'(a) - f'(b) < 0$ since $b - a > 0$. We have thus shown that:

$$\forall \alpha, b \in [x_1, x_2] : (a < b \Longrightarrow f'(a) < f'(b)),$$

and therefore $f' \uparrow [x_1, x_2]$. □

The forward part of the above proof is a straightforward calculation. However, the converse argument is non-trivial and cannot be found in most textbooks.

Altogether, since both Theorem 2.2 and Theorem 2.3 give logical equivalences, an immediate corollary is a statement establishing the equivalence of the strict secant line definition and the tangent line definition. This corollary does not require a separate proof and it reads:

Corollary 2.4: *Assume that f is differentiable at $[x_1, x_2]$. Then:*

$$\begin{aligned} &f \text{ convex up at } [x_1, x_2] &&\Longleftrightarrow \forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) > 0), \\ &f \text{ convex down at } [x_1, x_2] &&\Longleftrightarrow \forall x, x_0 \in [x_1, x_2] : (x \neq x_0 \Longrightarrow g(x|x_0) < 0). \end{aligned}$$

Finally, the second derivative test of convexity is also an immediate consequence of theorem 2.2 and can be established in merely two steps. For example, for the case of a convex up function, we have:

$$\begin{cases} f \text{ twice differentiable in } [x_1, x_2] \\ \forall x \in (x_1, x_2) : f''(x) > 0 \end{cases} \Longrightarrow f' \uparrow [x_1, x_2] \\ \Longrightarrow f \text{ convex up at } [x_1, x_2].$$

A similar argument can be given for the case of a convex down function. The first step uses the monotonicity theorem and the second step uses Theorem 2.2. However, whereas the second step can be reversed, the first step cannot be reversed without using a more general version of the monotonicity theorem. On the other hand, for most occasions the forward direction is sufficient.

3. Conclusion

Most textbooks do not give a complete development of the theory of convexity as detailed in this paper. This thorough treatment is imperative in analysis courses. For calculus courses, it may not seem reasonable to give all of the above proofs during lecture. However, we believe that instructors can still be careful enough to state the strict form of the secant line definition as the definitive definition of convexity, and then state without proof the lemma and the two theorems, all of which can be easily explained geometrically. Furthermore, instructors can still provide their students a copy of the proofs as a handout. There are always a handful of students in our classrooms with insatiable curiosity that will be truly delighted to see how these results are established.

Acknowledgements

It is a pleasure to thank Dr. Lokenath Debnath for discussions and for encouraging me to write this paper.

References

[1] J. Stewart *Essential Calculus. Early Transcedentals*, Thomson Brooks/Cole, Belmont CA, 2007.
[2] J. Haas, M. Weir, and G. Thomas *University Calculus. Early Transcendentals*, Pearson Education, Boston MA, 2012.
[3] D. Varberg, E. Purcell, and S. Rigdon *Calculus*, Pearson Prentice Hall, New Jersey NJ, 2007.
[4] S. Lang *A first course in Calculus*, Springer, New York NY, 1986.
[5] S. Lang *Undergraduate Analysis*, Springer, New York NY, 1997.
[6] T. Apostol *Calculus. I*, John Wiley & Sons, New York NY, 1967.